Place Recognition via 3D Modeling for Personal Activity Lifelog Using Wearable Camera

Hazem Wannous, Vladislavs Dovgalecs, Rémi Mégret, Mohamed Daoudi

Place Recognition via 3D Modeling for Personal Activity Lifelog Using Wearable Camera

Hazem Wannous[1,2], Vladislavs Dovgalecs[1],
Rémi Mégret[1], and Mohamed Daoudi[2]

[1] IMS, UMR 5218 CNRS, University of Bordeaux, Talence, France
[2] LIFL, UMR 8022, University of Lille, Telecom Lille 1, Villeneuve d'Ascq, France
{hazem.wannous,mohamed daoudi}@telecom-lille1.eu,
{vladislavs.dovgalecs,remi.megret}@ims-bordeaux.fr

Abstract. In this paper, a method for location recognition in a visual lifelog is presented. Its motivation is the detection of activity related places within an indoor environment to facilitate navigation in the lifelog. It takes advantage of a camera mounted on the shoulder, which is primarily designed for the behavioral analysis of Instrumental Activities of Daily Living (IADL). The proposed approach provides an automatic indexing of the content stream, based on the presence in specific 3D places related to instrumental activites. It relies on 3D models of the places of interest that are built thanks to a lightweight semi-supervised approach. Performance evaluation on real data show the potential of this approach compared to 2D only recognition.

1 Introduction

First person audio-visual sensing has recently emerged as a way to record the actions and location of a person within a visual life-log in order to provide data for activity and behaviour monitoring as well as memory aid. Our work is motivated by the application to medical behaviour diagnosis, where such technology enables medical practitioners to enrich their study of the behavior of subjects at home in their ecological environment. In particular, for early diagnosis of dementia is still a great challenge to prevent insecurity and health worsening in aged people living at home. Generally, diagnosis of possible dementia is asserted by comparing evidences, added to an autonomy decline. This autonomy decline is frequently assessed by Activities of Daily Living (ADL) interviews in which impairments are known to be related to cognitive decline caused by dementia. The approach developed in this article is related to the paradigm of the IMMED project, where wearable video is used to capture the Instrumental Activities of Daily Living (IADL) of a patient in detail, to be further reviewed or analyzed by medical experts [13].

To efficiently browse such data, content based indexing is required. Location is a powerful information that can be then used as input for activity inference. In [3], Conaire et al. analyzed the performance of 2D local feature matching for place recognition in visual lifelogs. Most of the existing work for visual lifelog

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 244–254, 2012.

indexing has indeed focused on 2D approaches using image classification based on global signatures such as Bag-of-Features [14,5], possibly refined by direct image to image matching based on local 2D features such as SIFT or SURF keypoints [9].

In this paper, we will evaluate the gain of using a 3D model of the places of interest for detecting events of interest within a visual lifelog captured by a wearable wide-angle camera. The method we propose instead leverages the reconstruction of 3D models of the places of interest using Structure from Motion (SfM) techniques and then apply 2D to 3D matching in order to estimate the accurate pose of the query image with respect to the 3D scene models. This approach requires a one-time bootstrap acquisition for the modeling of places of interest, which are in limited number in a home. The gain lies in the possibility that the estimated 3D localization can be further analyzed to extract activities related location events that are used to index the lifelog, in a more precise way than pure image recognition.

The paper is organized as follows. In section 2, we review works related to the location detection problem and position our solution. In section 4, the wearable camera setup and general architecture of the system will be introduced. In section 4, the methodology used for building the environment model will be explained. In section 5, this model will be used to localize the person and provide event detection. In section 6, the results of the proposed methods on representative situations will demonstrate the usefulness of the approach, and show the gain compared to frame based 2D recognition.

2 Related Works

Life-logging systems aims at performing the concept of digitally capturing daily activities and personal memories for later retrieval [18,3,5]. Doherty et al. [4] augmented streams of Lifelog images with geographic data by including locational information provided by a GPS unit rather than estimated from the visual content. Torre et al. [19] recorded the actions and displacements of several subjects using fixed cameras, motion capture, inertial sensors and head-worn narrow angle cameras, that can provide precise data on the movements and ongoing instrumental activites in a restricted place. We consider in our indoor context that the only data available comes from the wearable imaging device, which allows light-weight deployment, by not requiring to equip the house with smart-home sensors, such as cameras, presence sensors, radio-frequency RFID or Wi-Fi beacons. GPS is also not available, due to multipath and fading signals.

Under these constraints, Blighe and O'Connor [2] described a framework for recognizing real-world locations from passively captured images by the Microsoft SensCam sensor. They have classified each image scene into a number of event categories using image keypoints descriptors (SIFT), providing a simple tool allowing the user to annotate the image sets based on user events. Kang et al. [9] proposed to refine the local feature based recognition using Re-Search, to decrease ambiguity in environments with visual aliasing such as offices. Sundaram

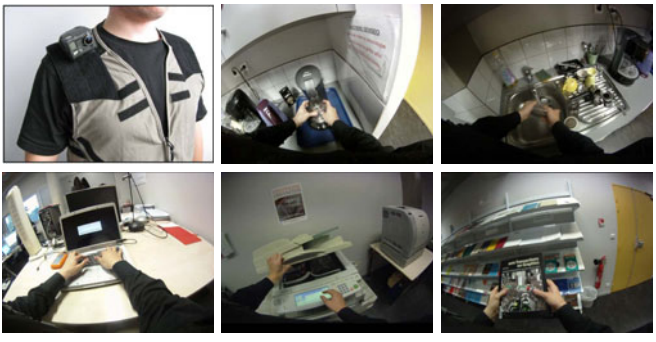


Fig. 1. Wearable camera and examples of places related to instrumental activities

et al. [18] extended image based lifelogging devices to the analysis of instrumental activity at video framerate. They infer location and manipulation activities using a Dynamic Bayesian Network, estimating location using the passage at doors and objects manipulations. Wearable video rate camera was also used by Dovgalecs and al. [5] who applied a Bag-of-Features approach to classify images according to predefined rooms. Kourogi et al. [10] use the fusion of 3D localization based on the alignement to geo-registered images of the environment and inertial sensors.

The work of Snavely et al. [17] have shown the possibility of estimating 3D models of scenes in a almost automatic way, thus enabling the creation of reference 3D models without the burden of geo-referencing the full environment. Its use for location recognition presents the advantage to necessitate only one model, instead of multiple images, as noted by Irschara et al. [8] who proposed outdoor localization based on 2D to 3D matching. Our work considers lifelogs taken with wearable video cameras, and evaluates the fitness of the 3D approach to extract place related events related to instrumental activities.

3 Overview of the System

3.1 Wearable Camera Setup

For a wearable camera to capture the instrumental activities with low motion, we opted for positioning the camera on the shoulder as in [18] and [13], integrated in a ergonomic clothing that is comfortable to the user. As the device is going to be used in capturing both the widest field of view of activities and the general context of the action, we chose to equip our prototype with a camera featuring a Fisheye lens with an effective diagonal angle of 150°and HD resolution (1280x960 pixels). This setup allows us to capture both the instrumental activities near the body and the environment as illustrated in Figure 1.

The captured images cover a very wide field of view, which help providing more varied matches for the image to image correspondences and thus improve

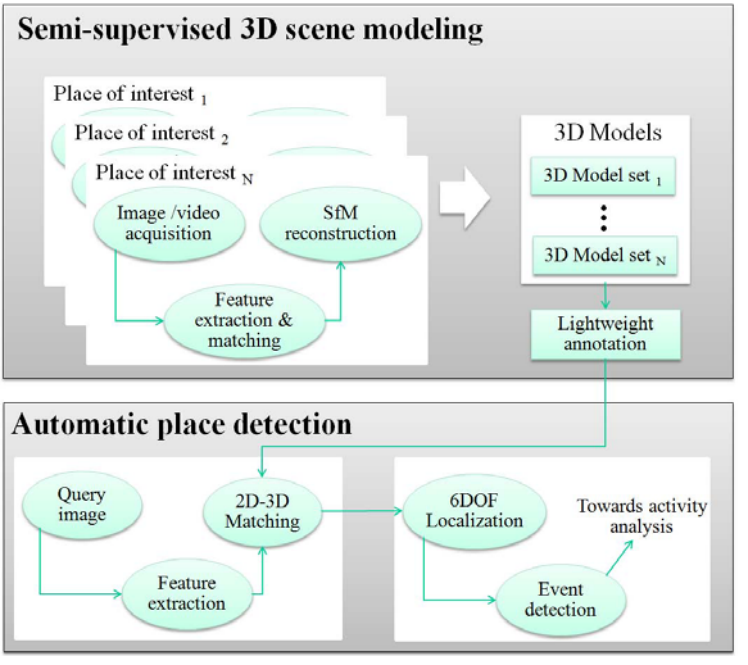


Fig. 2. Overview of the conceptual framework

the robustness of the SfM reconstruction. On the downside, radial distortion in such lens is usually very large and has to be taken into account. For this purpose, we calibrated the camera according to an omnidirectional camera model adapted for very wide angle lenses [15].

3.2 Global Architecture

The system architecture can be divided into two main modules: a semi-supervised 3D modeling of the places of interest and an automatic indexing of the lifelog (Figure 2). The first module deals with the recording of the image for 3D reconstructions and the generation of the 3D structures of the places of interest using SfM. Furthermore, we describe how to generate metric representation of the scenes and annotate the place of interest related to specific IADLs. The second module is composed of an automatic estimation of the 6 degrees of freedom (dof) with respect to the current place of interest, which is followed by an analysis in terms of events.

4 3D Scene Modeling

4.1 3D Reconstruction

We describe here the SfM-based reconstruction process. In orrder to have a lightweight protocol, the environment model has to be built using short training

sequences in a way that does not require complex manipulations while still providing good recognition rates in the lifelog. The analysis is focused on the places of interest (kitchen working plane, in front of the library, ...), in order to index the video with respect to the presence in predefined activity zones, which is suitable for a higher level analysis of the IADLs. Since we are interested mainly in some places of interest, and not in the complete environment, the reconstructions of each place is done independently.

Initially, our system assumes multiple training image for the reconstruction in the form of several 2D views of a place of interest taken from different angles and grabbed from video sequence. This is easily done by moving lateraly in front of the places of interest, which has to be done only once for each place, while wearing the camera device.

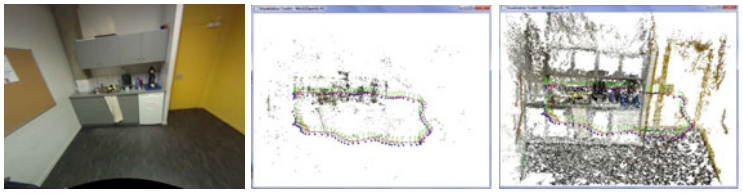
To extract the geometry between camera views, we need to match points between those views, which is possible using the SIFT local features [12] that exhibit good invariance properties. The captured images contain typically between 800 and 2000 SIFT keypoints. The SIFT features from image pairs are matched by considering the Euclidian distances between their 128-dimensional descriptors using the Approximate Nearest Neighbors / kd-tree package of [1]. Once an initial matching is obtained, the 3D reconstruction is done using the method proposed in [17]. The fundamental matrix for each pair images is robustly estimated using the RANSAC method [6]. During each RANSAC iteration, a candidate fundamental matrix is generated using the eight-point algorithm, followed by a non-linear refinement yielding a subset of geometrically consistent matches which will be chosen as an input to a SfM recovery algorithm in an incremental scheme. For each new camera, the back-projection error is minimized in order to calculate poses and 3D scene pointclouds by a generic Bundle Adjustment method, based on the sparse Levenberg-Marquardt algorithm [11].

Each model is represented as a sparse 3D point clouds with associated invariant visual descriptors. The initial training images are discarded in order to keep only a simplified representation of each place. Figure 3 shows the output of the SfM reconstruction applied on a sequence of 108 views captured from the kitchen place.

4.2 Annotation of the Places of Interest

Since the 3D point clouds obtained by SfM reconstruction are generally sparse, the scenes may be difficult to recognize directly of this model. To facilitate the annotation we use the Patch-based Multi-View Stereo method [7], which provides a denser 3D point cloud reconstruction from the sparser set of matched keypoints and estimated camera positions (see Figure 3). Although the sparser model is sufficient for the automatic estimation of location, the alignment between the two models being known.

The annotation of each place consists in defining the reference position in the related zone of interest; for example define the PC position in the office model, the working area in the kitchen, etc.



(a) (b) (c)

Fig. 3. Output of 3D reconstruction: (*a*) distorsion compensated image of the kitchen working area (*b*) sparse 3D model and camera poses obtained by SfM (108 cameras & 5741 points) (*c*) denser 3D point cloud obtained by PMVS (101198 points) for the interactive annotation of the environment.

First, a scale metric is applied manually to all reconstructed models using a graphical interface. Then, within each place model, the reference positionis directly selected in the 3D scene by clicking in the point cloud structure determining the center of gravity of the object. Another way to annotate an object consists in dragging a 2D box around a region of the current image containing this object; the center of the selected 3D points corresponds to the reference point of the corresponding place. Our approach has been applied to video capture, but it can be fully applicable to data captured at lower framerate, like those of [3], provided that the places of interest are captured from different points of view in order to reconstruct the reference 3D models.

5 Automatic Place Detection

5.1 3D Localization Using Natural Landmarks

Once the sparse 3D model has been annotated, it can be used in the automatic chain to index the video lifelog. SIFT keypoints are extracted from each query frame in the lifelog and compared to the models using robust 2D-3D matching, in a similar way as appending a new image to the 3D model [17]. Matching is first done on the SIFT descriptor features. The projective model is then used in the RANSAC framework for geometric verification. The estimation of the accurate pose of the query image with respect to the 3D scene models is then done iteratively. Tracking the natural landmarks of the environment by 2D-3D matching allows us to estimate a 6 dof trajectory in free movement of the person with respect to the reference models. These trajectories are then analyzed to detect events of interest.

5.2 Detection of Location Events

We aim to detect location events related to places that are meaningful with respect to the IADLs. The trigger of a location event of a place of interest depends primarily on the localization relative to the reference point associated to each place. If the current 2D frame image is matched with a 3D place model,

thresholding the distance between the camera and the predefined reference point delineates areas defined as: close (manipulation zone), intermediate (approaching) and far (seeing the place, but too far to do instrumental activities). Since the wearable camera is located on the shoulder, just above the arm associated with the dominant hand, we use the camera position directly to evaluate this distance, as we consider that the distance between the camera and a point in the environment is representative of the distance for manipulation. In the current state, only the presence in the close zone is considered to trigger a location based event, although information is available to define additional types of events. The frame based classification is segmented into event intervals using connected components of consecutive same class frames. Each frame is therefore associated to zero or one event. Each event corresponds to a temporal interval representing the arrival, stay and exit in a specific place.

6 Experimental Evaluation

6.1 Dataset and Methods

For the evaluation of our approach we used a lifelog of 36 minutes (68047 frames at 30fps). A volunteer weared the camera on the shoulder with free movements. The activities of the subject are organized around 6 places of interest: 1) sitting at the PC, 2) in front of the library, 3) using the copier, 4) in the lounge, 5) fetching items in the office closet, 6) preparing food in the kitchen. The lifelog was manually annotated for the events of interest with respect to these places, which corresponds to 42 temporal intervals of interest, separated by transition intervals with displacement or activities not related to the places of interest. Instances of these events are shown in Figure 1.

The 3D models of the places of interest were obtained by applying the SfM method, presented in section 4, to six very short sequences shot with the wearable camera and lasting 3 to 5 seconds each. An example of the result of our indexing method is shown with the associated groundtruth in Figure 4; it can be seen that the overall structure of the video with respect to place detection is correctly estimated, which we now evaluate more precisely.

We compared the proposed 3D method to the standard approach based on 2D correspondences [16,3]. It uses the same 2D-2D pairwise matching as in the 3D reconstruction step, with outlier exclusion using the RANSAC procedure applied to fundamental matrix estimation. The temporal majority vote filtering was also applied on this approach. In the 2D approach, a query frame is matched to the class of the reference images, used in the construction of reference 3D models, for which it has the largest number of inlier matches. When the number of inliers is below a threshold of 100 inliers (best parameter found empirically), the frame is considered rejected.

6.2 Evaluation

The evaluation was done both at frame and event levels. The output of the location recognition module is an estimated location vector f that maps each

Table 1. Performances of 2D and 3D approaches for place recognition. Temporal window 5s.

	2D method		3D method	
Framewise global *Accuracy*	0.368		0.627	
Framewise class based $Prec_i$, $Recall_i$	Precision	Recall	Precision	Recall
1 - PC	0.018	0.232	0.403	0.934
2 - library	0.453	0.522	0.799	0.613
3 - copier	0.338	0.723	0.682	0.907
4 - lounge	0.053	0.038	0.682	0.832
5 - office	0.066	0.270	0.877	0.511
6 - kitchen	0.333	0.687	0.575	0.928
Framewise global *Prec* /*Recall*	0.451	0.2168	0.865	0.575
Eventwise global $Prec^{evt}$ / $Recall^{evt}$	0.520	0.190	0.814	0.523

frame $n \in \{1, \ldots, N\}$ either to a class of interest $f(n) > 0$ or to the reject class $f(n) = 0$. We also denote by $g(n)$ the corresponding ground truth location. Each frame can belong to one of the following cases; correct classification can be either a True Positive (TP : $f(n) = g(n) > 0$) or True Negative (TN : $f(n) = g(n) = 0$); incorrect classification can be either False Negative (FN : $f(n) = 0,\ g(n) > 0$), False Positive (FP : $f(n) > 0,\ g(n) = 0$), or, due to multiclass classification, Incorrect Positive (IP : $f(n) > 0,\ g(n) > 0,\ f(n) \neq g(n)$). The global accuracy metric is defined as the ratio of true positive frames to the total number of frames $Accuracy = \#\text{TP}/\text{N}$. Since we are dealing with data that contain a large amount of reject class, we also use a multiclass global precision defined as $Prec = \#TP/(\#TP + \#FP + \#IP)$, and recall defined as $Recall = \#TP/(\#TP + \#FN + \#IP)$, Class specific precision $Prec_i$ is defined by considering only frames such that $f(n) = i$, class specific recall $Recall_i$ is defined by considering only frames such that $g(n) = i$.

An event E is defined as an interval of consecutive frames associated to the same class. An estimated event is considered correct if it overlaps on more than 50% of its length groundtruth events of the same class; a groundtruth event is considered retrieved if it overlaps on more than 50% of its length estimated events of the same class. Each estimated event can therefore be evaluated as True Positive or False Positive, yielding event based Precision $Prec^{evt} = \#TP^{est\_evt}/(\#TP^{est\_evt} + \#FP^{est\_evt})$. Each groundtruth event can be either True Positive or False Negative, yielding event based Recall $Recall^{evt} = \#TP^{gt\_evt}/(\#TP^{gt\_evt} + \#FN^{gt\_evt})$.Performance in place recognition was assessed on the precision, recall and accuracy rates at frame level, and precision and recall at event level (see Figure 5). The confusion matrix associated to the rate of correct detection for each place is given in Figure 6.

On all measured metrics, the proposed 3D approach outperforms the 2D approach. The additional constraints and completeness stemming from the use of

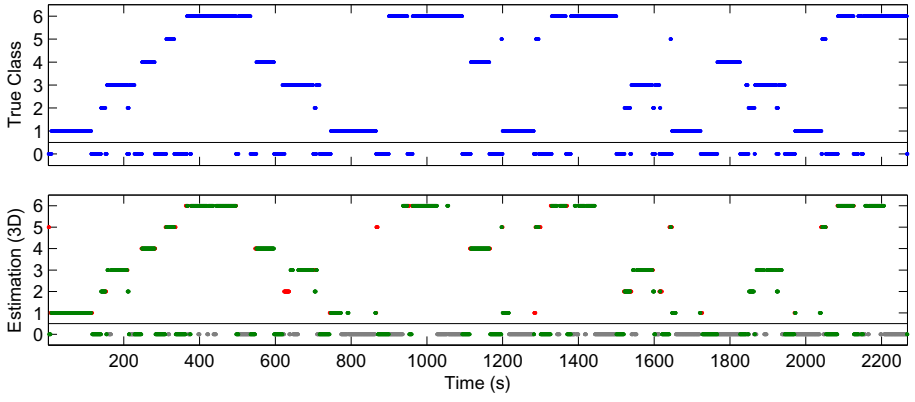


Fig. 4. Chronogram on the test lifelog. Top: true place. Bottom: estimated place using the proposed method (temporal window of 5s). Color-code: green=correct, red=incorrect, gray=false negative.

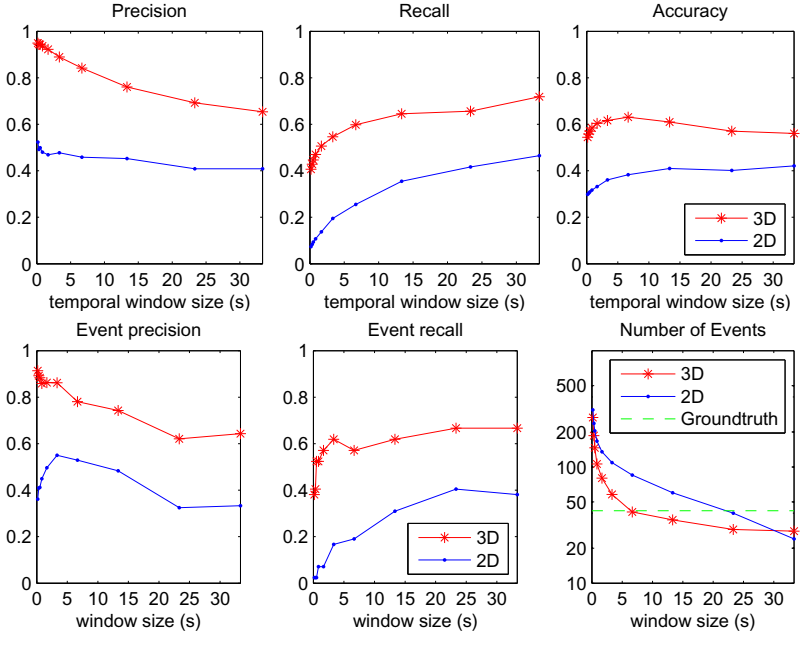


Fig. 5. Evolution of Precision, Recall, Accuracy and Number of Events as a function of the temporal window size. (Top) Frame-based metrics (Bottom) Event-based metrics.

3D models merging information from several reference images therefore help exclude cases that are incorrectly accepted by the 2D approach while providing a better recall that using a separate comparison with each reference frame. This is observed on classwise as well as global metrics (Table 1). This improvement is

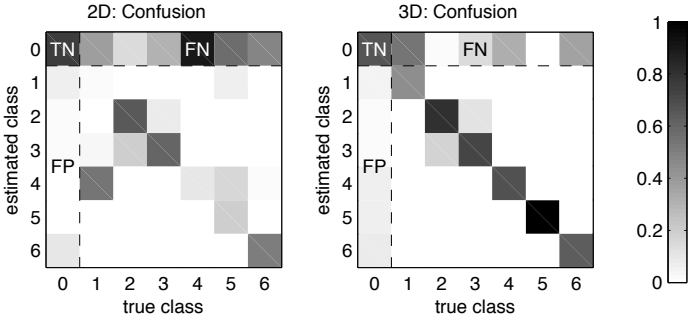


Fig. 6. Frame-based confusion matrix of place recognition for (left) 2D and (right) 3D based matching, both with temporal window of 5s

stable with respect to the choice of the temporal window (Figure 5). In particular, we can notice a sharp increase of the event recall for short-term temporal windows (Figure 5, bottom-center) which shows that some scattered false negatives, probably due to temporary occlusion or motion blur can be compensated. A longer temporal window is required when using the 2D approach. These observations can also be related to the lower number of events detected by the 3D approach (Figure 5, bottom-right), that corresponds to less oversegmentation. This allows the use of shorter temporal windows for postprocessing, thus generating less artifacts due to the temporal regularization.

7 Conclusion

In this paper, we have addressed the need of providing event based place detection for behavior monitoring in visual lifelog, by exploiting the same camera sensor that is used for observing the instrumental activities: a wearable camera fixed on the person shoulder. The proposed system is based on 3D models obtained using SfM techniques. It was shown suitable for the detection of situations of interest for the analysis of IADL, and improves the false alarm rate compared to purely visual recognition based approaches. The current vision only system is dedicated to video monitoring in an indoor home environment and showed promising results. We intend to expand the experimental part to a larger set of users, and extend the approach by hybridizing with inertial sensors in order to improve the accuracy of the localization and its robustness to occlusion and motion blur, which is the subject of future work.

Acknowledgement. This work is partly supported by a grant from the ANR (Agence Nationale de la Recherche) with reference ANR-09-BLAN-0165-02, within the IMMED project. It was initiated during the stay of M. Hazem Wannous at IMS.

References

1. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y.: An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. Journal of the ACM 45(6), 891–923 (1998)
2. Blighe, M., O'Connor, N.: Myplaces: Detecting important settings in a visual diary. In: ACM International Conference on Image and Video Retrieval, Niagara Falls, Canada, July 7-9 (2008)
3. Conaire, C.Ó., Blighe, M., O'Connor, N.E.: SenseCam Image Localisation Using Hierarchical SURF Trees. In: Huet, B., Smeaton, A., Mayer-Patel, K., Avrithis, Y. (eds.) MMM 2009. LNCS, vol. 5371, pp. 15–26. Springer, Heidelberg (2009)
4. Doherty, A., O'Conaire, C., Blighe, M., Smeaton, A., O'Connor, N.: Combining image descriptors to effectively retrieve events from visual lifelogs. In: ACM Multimedia Information Retrievaly, Vancouver, Canada, October 30-31 (2008)
5. Dovgalecs, V., Mégret, R., Wannous, H., Berthoumieu, Y.: Semi-supervised learning for location recognition from wearable video. In: CBMI (2010)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 381–395 (1981)
7. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Trans. on PAMI 32, 1362–1376 (2010)
8. Irschara, A., Zach, C., Frahm, J.-M., Bischof, H.: From structure-from-motion point clouds to fast location recognition. In: CVPR, pp. 2599–2606 (2010)
9. Kang, H., Efros, A., Hebert, M., Kanade, T.: Image matching in large scale indoor environment. In: First Workshop on Egocentric Vision (2009)
10. Kourogi, M., Kurata, T.: A method of personal positioning based on sensor data fusion of wearable camera and self-contained sensors. In: IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems, pp. 287–292 (2003)
11. Lourakis, M., Argyros, A.: The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. Technical Report 340, Inst. of Computer Science-FORTH, Heraklion, Crete, Greece (2004), `www.ics.forth.gr/~lourakis/sba`
12. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
13. Mégret, R., Dovgalecs, V., Wannous, H., Karaman, S., Benois-Pineau, J., El Khoury, E., Pinquier, J., Joly, P., André-Obrecht, R., Gaëstel, Y., Dartigues, J.: The IMMED project: wearable video monitoring of people with age dementia. In: ACM Multimedia, Firenze, Italy, pp. 1299–1302 (2010)
14. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR, vol. 2, pp. 2161–2168 (2006)
15. Scaramuzza, D., Martinelli, A., Siegwart, R.: A flexible technique for accurate omnidirectional camera calibration and structure from motion. In: ICVS (2006)
16. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV, vol. 2, pp. 1470–1477 (2003)
17. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. IJCV 80(2), 189–210 (2007)
18. Sundaram, S., Mayol-Cuevas, W.: High level activity recognition using low resolution wearable vision. In: First Workshop on Egocentric Vision (2009)
19. Torre, F., Hodgins, J., Montano, J., Valcarcel, S.: Detailed human data acquisition of kitchen activities: the cmu-multimodal activity database. In: HCI Workshop (2009)